

Work Projected presented in the context of a Directed Research Internship at the Directorate-General of Statistics of the Portuguese Ministry of Education, and as part of the requirements for the Award of a Masters Degree in Economics from the NOVA - School of Business and Economics

# Predictors of student success: a study of Portuguese Higher Education graduates

José Miguel Lourenço Cerdeira  
Student number 628

This project was carried out on the Public Policy Major, with the supervision of:  
Ana Balcão Reis, Maria do Carmo Seabra and Luís Catela Nunes

January 7th, 2015

# Predictors of student success: a study of Portuguese Higher Education graduates

## Abstract

This research analyses the determinants of academic achievement of Bachelor and Integrated Masters students that graduated in 2011/12 in Portugal. It uses student individual data on student's characteristics, social background, and past educational success to predict deviation from average scores and whether or not they graduate in due time. The main findings are that parent's education effects are totally factored in basic and secondary education (although affecting Higher Education performance indirectly) and that internal high school scores are better predictors of success at the university level than National Exams. Moreover, the findings show that the level of significance and relevance of factors like working status, social support and gender, vary with the type of Higher Education (University versus Polytechnic and more demanding versus less demanding degrees).

**Acknowledgements:** I must thank Professors Ana Balcão Reis, Maria do Carmo Seabra and Luís Catela Nunes for all the advice and precious help. I would also like to thank everyone at the Ministry of Education for their support, in particular Dr. Carlos Malaca for the motivation he gave me, Professor João Baptista for his contributions, and Professor Luísa Loura for her valuable opinion, material and institutional support. Most importantly, I thank my parents, to whom I dedicate this piece of work.

# 1 Introduction

Today, in Portugal, over 28% of the population between 25 and 34 years old have completed some form of Higher Education.<sup>1</sup> This number is lower than in many developed countries, but the reality is improving. In general, for the economy, most of the literature reveals that there are benefits to be reaped from this qualification, both individually and for countries<sup>2</sup>. These gains are explained in economic theory by the model of Human Capital proposed by Becker in 1964<sup>3</sup>. In this model, the investment in education is similar to any other kind of investment, except for the fact that human capital can only be “rented” by the person who invested in it, not being possible to sell this capital to other economic agents. Moreover, this kind of investment is not completely certain nor uniform. Contrarily to a monetary investment in a machine, in which the return does not vary much, two people may do the same nominal investment in human capital (for example, finishing High School) and reap completely different benefits from it. In other terms, this investment may be more or less successful.

Still, what can one consider a success in the investment in human capital? Ultimately, this can only be measured by the future benefits in the economic activity of the worker, whether these are in terms of his salary, his well-being, or in measures related to his social inclusion. Although these measures are not easily associated and found in datasets together with academic records, there are simple and effective proxies, given by the academic process in itself: the academic achievement of the student. Basically, one can evaluate whether the investment in the education of an individual is successful or not

---

<sup>1</sup>“Education at a Glance 2014”, OECD.

<sup>2</sup>Krueger and Lindahl (2000) show that “the effect of changes in educational attainment on income growth in cross-country data is at least as great as microeconomic estimates of the rate of return to years of schooling”.

<sup>3</sup>Gary Becker proposed this model in his book, Human Capital. This is not the only theory regarding the economic value of education. There is another model, signalling, that states that productivity of individuals is innate, determined by its ability, and that education is a method for the market to evaluate that same ability, and for workers to signal that they do possess higher ability. These models are not necessarily exclusive, and, it may be that, in many levels of education, in different combinations, there is part of investment in human capital, and part of signalling innate ability.

by checking if the goals of such investment were reached. To do so, one must analyse if the qualification was completed and, if so, with what quality, evaluated by scores. This has been extensively done in some levels of schooling, and, to some extent, in the area of Higher Education. In this study, we propose to look at this issue through the use of a national database of graduates in the year 2012, from the Directorate-General of Statistics for Education and Science (from the Portuguese Ministry of Education and Science), joined with other databases with individual microdata of students.

In particular, this study tries to answer a simple question: what determines the success of students in Higher Education? To answer this, ideally, it would be necessary to possess data on a cohort entering Higher Education, with knowledge of what happened to these students, if they dropped out, transferred, or graduated, and if so, in how long and with what final score. However, the data available is not as comprehensive<sup>4</sup>. Thus, this study will focus on three specific questions that help to give insight to the issue just mentioned. **1. Conditional on the fact that students are enrolled in a particular degree/class and succeed in graduating, what are the predictors of the deviation of their final score from the average final score in that class?** **2. Conditional on the previously referred fact, which factors predict whether the students finish their degree later than the standard duration of the qualification?** The conditioning on entry in a particular degree is necessary because students are not distributed at random between schools and degrees. The best students go to departments with better reputation, which in turn are normally seen as more demanding. Thus, even deviation from average score (or the probability of having a delay in graduation) is not comparable among institutions. Therefore, it seems natural to ask what determines this distribution of students. The third question relates to this

---

<sup>4</sup>In particular, as we only have data for students graduated in 2012, we cannot ask about the determinants of graduation.

problem. As in Portugal entry is mostly determined by the mix between National Exam scores and High School grades, we use these data to obtain some insight into what may be the reasons for the existing distribution. We ask, then: **3. What are the best predictors of the Higher Education Access Score and its components?**

This research and following discussion is important for the field of Economics of Education, but also for Economics in general. It is important to know what determines the success of students in Higher Education, both for the information of current and future students and parents, but in particular for policy makers. With this kind of information, it is possible, for example, by realizing which factors influence the probability of a delayed degree, to check the efficiency (evaluated in terms of the extra years needed for a student to finish a degree) of a particular university's degree, taking into account the profile of the students, according to the identified influencing determinants. Although data about Higher Education is scarce, in lower levels, Portugal is one of the least efficient countries in the OECD in using the resources in education.<sup>5</sup> Thus, facing a future horizon of public spending cuts, it can prove extremely helpful to compare institutions and promote reforms in those organizations where these problems are most relevant. Moreover, in terms of equity, it can be relevant to know how much of a student's success is determined by these variables, so that policies aiming at a larger student achievement are actually targeted at the issues that are effectively important.

## 2 Literature Review

The field of Economics of Education is a fairly recent one, although its research is now numerous and diverse. It owes much to the contributions of Jacob Mincer, Gary

---

<sup>5</sup>In "The Efficiency Index", GEMS Education Solutions identifies Portugal as the 24th out of 30 countries that submit their students to the PISA tests, in terms of efficiency of the resources spent in basic and secondary education, in the last 15 years. For an exploratory study of the efficiency question on Higher Education in Portugal, see Cunha and Rocha, "On the Efficiency of Public Higher Education Institutions in Portugal."

Becker and Theodore Schultz with the development of, respectively, the theories of the Returns to Education, the Human Capital Model and the latter highlighting the “Economic Value of Education”. As to the discussion of which are the determinants of student performance, it started with the Coleman Report, which shed some light on the relevance of various inputs in academic success. After this, in the 70’s, Hanushek proposed a formulation of these relationships using the Education Production Function, which is much similar, in form, to any particular firm’s production function (Hanushek, 1979). Since then, many different researchers have looked at the determinants of success in education. One of the most important determinants has been found to be the family background of the student, commonly evaluated through the parents’ level of education: Simkins (2005) finds that “what matters much more than gender, urban/rural location and category of household is the education of the household head”. Plug and Vijverberg (2001) state also that “children raised in highly educated families are more educated than children raised in less educated families”. However, it is not completely clear how this relationship works. Feinstein and Symons (1999) look at this effect through a teacher-reported variable called “parental interest” and claim that “the major influence on attainment is parental interest (...). This dwarfs the direct effects of parental education and class, but is itself strongly correlated with these.”

More recently, studies on the determinants of success in education have been applied to Higher Education. In the UK, Smith and Naylor (2001) have looked at the determinants of “degree performance” and found several significant factors. Besides confirming that “academic performance at university is better the more ‘advantaged’ is the student’s home background”, they report that performance at the A-levels, attendance of a Local Education Authority School (in comparison with “independent” schools), living at ‘home’ (not being displaced), being married, studying at full-time, and being female

are all characteristics that increase one's predicted performance. Horowitz and Spector (2005) confirm these results for national high school exams, gender, and the effect of private schools, although they distinguish private schools from religious private schools, which in turn seem to have a positive effect on performance. This effect was also reported on the entry to college by Evans and Schwab (1995). In regard to the effect of school resources, McNabb et al. (2002) state that, although "higher staff-student ratio and library expenditure (per student) are found to increase student performance", "higher total expenditure per student does not necessarily enhance academic achievement". Looking at social support, Marcenaro and Navarro (2007) have found a positive influence of social scholarships in academic achievement. Studies on the time-to-degree are not as frequent. Still, Desjardins et al. (2002) find a relevant positive effect of High School GPA on timely graduation. Amann (2005) looks at the effect of employment on completion and time-to-degree, finding that both full-time and part-time employment have a negative influence on these variables. Other studies are more focused, like Siegfried and Stock (2001), on Ph.D. time-to-degree.

In Portugal, this type of study was only done at university level. The Universidade do Porto used its internal data on the characteristics of students to compare proportions of students with given characteristics and check if they were under/overrepresented in the group of students with best scores or in the group of students that did not fail subjects (Universidade do Porto, 2011). In their results, being female and having a greater Access Score (determined by High School grades and National Exam scores) seem to be correlated with academic success. Also, being 18 when starting Higher Education also appears to be correlated with better performance. As for the type of high school attended, the results do not seem to show any particularly significant correlation. Finally, Alves (2014) used data from a cohort of students at Nova School

of Business and Economics and his main findings were in line with the literature: High School grades and National Exam scores are both good predictors of the final GPA, with high-school grades having a much stronger predictive power; being 18 years old or less is also associated with larger success, as well as not being displaced. Moreover, in this case of the study of Economics or Management, “male students perform better”.

### **3 Data**

#### **3.1 Data analysis and database construction**

In the course of my internship at the DGEEC, dealing with data from different sources has been the main task. The final purpose was to pursue this research analysis. However, in order to accomplish that, various steps were necessary to build the database that was ultimately used. For a more detailed explanation of all the work done previous to this research, see the Appendix. In this section, some of the specific variable analysis pursued is presented below. One of the variables which were analyzed was the “number of registrations”. This variable existed in two different sources: in the information table of “Graduates” and in the table of “Registrations”. In the “Graduates” table, it was possible to find the number of registrations necessary to complete the degree. In the “Registrations” table, the numbers of registrations until (and excluding) the present registration were represented. Moreover, there were some specific concerns as to the reliability of the reported data. Thus, an evaluation of the data was done, in order to check the consistency of the information between the two sources. First, a simple analysis of the distribution of this variable on the “Graduates” table was made, regarding only students that had obtained a Bachelor degree in 2012. This distribution seemed reasonable, with a peak at 3 registrations and a progressive decrease from that maximum on. There were few observations with 1 or 2 registrations, and those were



justified by cases of students who transferred from other institutions. In addition, as both tables referred to the same school year of 2011/12, all individuals found in the same degree and institution would have to have matching values in these variables. For instance, if the student had 2 previous registrations in the “Registrations” table, that should correspond to a number of 3 registrations in the “Graduates” table, because being in the first table meant obviously that the student was registering for the third time.<sup>6</sup> Looking at all the students that completed a Bachelor degree in 2012, almost 20% of these had inconsistent values, i.e., values that were not matching between the two sources of information. Thus, for the matter of our research, only observations in which these two sources of information were consistent were used. Regarding the internal use of this assessment, the goal was for the DGEEC to be able to prevent this inconsistency, when receiving the reported data; and, at the same time, to demand more precision and explanations from entities that continuously deliver unreliable data. Two other variables were the subject of a specific report: ECTS credits and Parents’ level of education. Regarding the first one, an analysis of consistency between data available from common students in different years was conducted. The conclusion was that the reporting of this variable was completely unreliable<sup>7</sup>. This, in turn, led us not to use this measure as a dependent variable on our study. As for the second variable, a report was produced on its incompleteness. It was found that, for more than 20% of the cases, both parents had their schooling reported as “Unknown”. Moreover, an analysis of this part of the population indicated that this characteristic (having Parents’ Schooling as unknown) is correlated with lower scores and more time to finish the degree. Thus, it is probably not randomly present in all actual sublevels of schooling. Finally, another

---

<sup>6</sup>See Appendix B for a more detailed explanation of this process.

<sup>7</sup>Using these common students in subsequent years, it was found that some of the data were completely unrealistic. In some cases (and with very significant percentages), students actually lost credits from one year to another, something which is not possible in the system. Moreover, in some institutions, very large proportions of students (in many cases, more than 30%) had an addition of credits larger than possible by regulations.

document describing other minor issues with the rest of the database variables was elaborated.

Concerning the creation of the database used for the purpose of our research, its final content came from five different sources. The two main sources of information are the table of the “Registrations” (11/12) and the table of the “Graduates” (11/12). These have, respectively, individual information on registrations on a particular year, and individual data on completion of degrees in a specific year. This information was completed with a constructed “Exams” (2006-2012) table, with individual information from the National Exams and High School grades of thousands of students, from a National Exams database. Additionally, there was a “Degrees” table, also constructed by the author from different sources of information, with characteristics of the degrees such as the Area of Study where they are inserted, their duration, the number of students registered in 2011/12, etc. Finally, the Minimum Entry grades in 2013 for public Higher Education institutions were added to this database from publicly available data.

The first significant step in connecting these tables was to correct the identification numbers such that they were all represented in the same way: i.e., with 8 digits, transforming a “56789” into a “00056789”, in order for all the possible connections to be made. This was done both in the “Registrations” and the “Graduates” tables. The second step was to “cross” these two tables, connecting them into one, where individuals were the same, in the same establishment and degree, and, more importantly, imposing the consistency condition related to the “number of registrations” variable. Then this table was joined together with the “Degrees” table. For the use of the “Exams” table, various changes had to be made, so that the table had only the exams that actually

determined entry for Higher Education<sup>8</sup>, thus eliminating exams from previous years, selecting only the best exam if the student had done the “2nd phase” (repetition phase) of the exams, and not losing any observations at the same time. Then, to join the table containing the rest of the information, another condition was imposed: the year 2012 minus the number of registrations taken to complete the degree would have to be equal to the year in which the exams were taken. In this way, cases where students have asked for transference from one degree to another were eliminated, while making sure that the “number of registrations” coincided with “number of years taken to complete the degree”. Finally, after joining this information together in one table, the identification variables were dropped, so as to guarantee an anonymous database, adequate for research.

### 3.2 Final database

The final database is part of the population of students that finished a Bachelor or an Integrated Masters in the year of 2012.<sup>9</sup> Moreover, these are only students that came directly from High School, not making any interruptions, and that were not transferred from other Higher Education institutions. Also, due to lack of data on earlier years, only students that did their National Exams from 2006 on are included. Furthermore, there is no data on students that have not finished their degree (that either dropped out, or are still doing their academic path). Finally, as said during the description of the construction of the database, only students with fully reliable data (especially regarding the number of registrations) were considered. In total there are 27,412 observations. Of

---

<sup>8</sup>Still, there are cases where exams that actually determined entry were the ones in the previous years, as students can save their previous year’s score for applying to Higher Education. Moreover, the variable used, includes not only the exams used for entering in college, but all the other National Exams done in that year.

<sup>9</sup>Many Higher Education institutions in Portugal, after the Bologna process, turned their previous “Licenciaturas” into Integrated Masters, which maintained the 5 year structure corresponding to both the 1st and 2nd cycle of Bologna studies. In specific areas (like Engineering, Architecture, Medicine), this is the most common case. Thus, it is prudent to include these degrees in the analysis. For a more detailed description of the Portuguese Higher Education System, see Appendix C.

these, 23,632 are Bachelor students (out of a universe of 44,758 graduates in 2012) and 3,780 are from a Masters programme (out of 7,797). The information available includes:

- 1) Success measures – Final Score of the degree; number of years taken to graduate; and whether the degree was finished on time or delayed;
- 2) Previous academic information – High School GPA; Mean Score of the National Exams; Type of high school attended; Type of curriculum pursued at secondary school;
- 3) Socio-economic factors – gender; age; schooling of parents; if the student was displaced; if she/he received social support; and working status;
- 4) Institution variables – average Final Score in that degree in that year; average years taken to finish degree; whether the institution is public or private; and if it is a university or a polytechnic.

Let us now look at the descriptive statistics of our population.

Table 1: Summary statistics

Variable	Description	Obs	Mean	Std. Dev.	Min	Max
Bsc	Bachelor = 1, Master = 0	27412	.862	.345	0	1
FinalScore	Final Degree Score	27412	13.793	1.444	11	20
YearsTaken	Time taken to graduate	27412	3.987	.983	1	6
DelayedDegree	Graduated later than expected	27412	.348	.476	0	1
HSGrade	High School Final Grade (0-200)	27412	148.291	18.418	100	200
EMScore	Mean of the National Exams Score (0-200)	27412	127.333	27.502	10	200
PubHS	Attended public high school (only)	27412	.844	.363	0	1
BothHS	Attended both private and public high schools	27412	.045	.208	0	1
Academic	Academic track in high school	27412	.635	.481	0	1
Male	Male = 1, Female = 0	27412	.347	.476	0	1
Age	Age at entry	27412	19.007	2.114	17	53
PSchYrs	Average schooling of parents in years	20051	10.066	4.485	0	21
Displaced	Student is away from permanent residence	27412	.268	.443	0	1
SocialSupport	Student had social support scholarship	27412	.073	.261	0	1
Working	Student works while studying	27412	.05	.218	0	1
Public	Public = 1, Private = 0 (Higher Education)	27412	.791	.406	0	1
Univ	University = 1, Polytechnic = 0	27412	.547	.498	0	1

The sample is constituted by 86% graduates from Bachelor and 14% graduates from an Integrated Masters. The average final Score is about 14 out of 20, and 35% of the students have graduated later than they should have. The mean scores at, respectively, High School and the National Exams taken in the year before entering college were 14.8

and 12.7 out of 20. Most of these students came from public high schools, with 10% coming from private schools (85% from public schools, while 5% studied in both, during their High School years). Moreover, 63% of the students came from the main curricular path. The rest came from other high school curricula, that includes vocational and professional tracks. 35% of the graduates are male, with 19 as average age when entering college. About 46% of the students had parents who went to Higher Education, and the average schooling years of their parents is around 10 years. 26% of these students were displaced from their permanent residence (i.e. not living with their parents), 7% receive a Social Support scholarship and 5% of them were working at the same time as they studied. Finally, regarding the Higher Education institution where they graduated from, the sample has 79% of students from public institutions and 55% from universities (45% come from polytechnic institutions).

Comparing with the statistics for the population of 2012 graduates, the great majority is very similar, with the exception for the age categories and the working-student percentage. Considering Bachelor graduates, the population over 25 (when graduating) represents 38% in the population, while it is only about 12% in the sample. Also, while in the population of Bachelor students there are 12% working students, only 5.5% of those in the sample have this characteristic. This happens due to the fact that, in the construction of the database, we have only selected those students who came directly from their National Exams. From those students, only the ones that took their exams from 2006 on were considered. This excludes a significant portion of the student population: the ones registered in university with the “over-23” scheme (around 12% of the total registered students)<sup>10</sup>, and students that did exams prior to that. Also, in this

---

<sup>10</sup>The “Over-23” entering method allows that adults that are over 23 years old can apply for a Higher Education Institution, not having to present results from National Exams. Individual Institutions then establish methods of evaluation of their own to select the students, that may composed of written tests, oral interviews, or others.

final dataset, around 30% of the observations had missing values relative to parents' schooling level. Thus, multiple imputation was used to obtain these values.<sup>11</sup> Consequently, we have to consider that our results may only be applicable to this particular section of the population, taking this into account when analysing the empirical results.

## 4 Method

As described by various authors, education is a cumulative process. Thus, the optimum way to perfectly identify the effects of these variables on success would require having, not only the contemporaneous inputs, but also the historical data on these inputs, together with a measure of the innate ability endowment of each student. As this is not possible, we will use a Value-Added specification<sup>12</sup>, which considers contemporaneous inputs, and a lagged success measure as an instrument for past values of these same family inputs, school inputs, and innate ability.

$$y_{i0} = x'_{i0}\beta_1 + x'_{i-1}\beta_2 + \dots + \mu_i\beta_3 + \varepsilon_{i0} \quad (1)$$

$$y_{i1} = x'_{i1}\delta_1 + x'_{i0}\delta_2 + \dots + \mu_i\delta_3 + \varepsilon_{i1} \quad (2)$$

In the equations above,  $y_{i0}$  is the historical data on achievement (High School grades and Exam Mean scores, in this case), and  $y_{i1}$  is the variable that indicates success in Higher Education. In this specific case, in order to accomodate for different grading policies and different student bodies in each department, we use the deviation from the class Final Score mean. The characteristics and inputs, whether these are school-related, individual, or from family background, are represented by  $x_{it}$ , while  $\mu_i$  represents innate ability.  $\varepsilon_{it}$  represents the residuals, considered to be zero-mean i.i.d. and uncorrelated

---

<sup>11</sup>See Appendix D for a detailed explanation.

<sup>12</sup>Todd and Wolpin (2003) provide an accurate description of the various types of models possible for applying to regressions, considering the cumulative aspect of Education and taking into account circumstances like the availability of data.

with the independent variables.

We do not dispose of all the historical data on the inputs, but we do dispose of the past achievement result. If we combine both equations, we have an expression that considers contemporaneous inputs and this lagged success measure. Note that we are assuming that the effects of past inputs, that we do not include in our specification, are only felt in the past result  $y_{i0}$

$$y_{i1} = x'_{i1}\delta_1 + y'_{i0}\delta_2 + \varepsilon_{i0} \quad (3)$$

Although this solves the problem, we may still have issues of bias if there is measurement error in the previous success measure. Also bias can also arise if there are past inputs that do not influence the previous test scores, but do influence current success and are correlated with other inputs actually included in the specification.

As said previously, the first model will be based in the deviation of the Final Score from the Class average, *DevFinalScore*, as the dependent variable. For the estimation of this model, we will use Ordinary Least Squares.

The second model to be used is a Probit model to predict the probability of a student graduating later than it was supposed, according to the established duration of the degree. This is represented by the binary variable *DelayedDegree*, which assumes the value of 1 when the student finished the degree later, and 0 otherwise. Both these models reflect the prediction capability of the independent variables used, conditional on the fact that the student has, in fact, graduated, and was studying the Degree  $d$  in the Institution  $i$ .

Thus, and because of the fact aforementioned, it is reasonable to make an attempt on finding out the determinants of the distribution of students by degrees and institutions. Firstly, there is a matter of choice of the area of study. Secondly, and however, within that area of study, students tend to choose the institution/degree with the best rep-

utation. Institutions then proceed to selection, meaning that, on average, the better prepared students are allocated to the best institutions and degrees. The quality of students is evaluated mostly by their High School Grade and the score at specifically defined National Exams for each degree in an institution. More particularly, in the context of the Public Universities, there is a national procedure where students are ordered according to their scores<sup>13</sup>. Combining this ordering with the choices of students, these are then allocated to the universities. This means that, at least in the public sector of Higher Education, the High School scores and the National Exam Scores are a signal for this distribution. We may then use OLS to assess what are the determinants of such distribution by looking at the determinants of these two variables.

## 5 Empirical results and analysis

In this section, the results of the econometric models are presented. The first model uses Ordinary Least Squares (with Multiple Imputation when using the variable related to Parent’s Schooling) to look at the predictors of the deviation from the average Final Score of the class, for each student, *DevFinalScore*<sup>14</sup>. The independent variables considered are the following: *EMScore*, *HSGrade*, *Male*, *Working*, *Support*, *Displaced*, *PubHS*, *BothHS*, *Academic*, *PSchYrs*. The equation estimated by OLS follows:

$$DevFinalScore = \alpha + \beta_1 EMScore + \beta_2 HSGrade + \beta_3 Male + \beta_4 Working + \beta_5 Support + \beta_6 Displaced + \beta_7 PubHS + \beta_8 BothHS + \beta_9 Academic + \beta_{10} PSchYrs + \varepsilon$$

Yet, in order to get robust results, various specifications are considered: i) using only

---

<sup>13</sup>After finishing High School and National Exams, students submit an application with six ordered preferences of their degree and institution of choice. For each degree, in an institution, there is at least one specific National Exam to be presented in such application. The combined score (usually 50/50) of the “specific subject” National Exam and the High School Final grade determine an Access Score. Combining the preferences of the options with the particular listings of ordered students for each degree, in each department, students are then allocated by that order to the Public Higher Education Institutions. For more detail, see Appendix C.

<sup>14</sup>Final Scores of a completed degree vary between 10 and 20.



*EMScore* as predictor; ii) using only *HSGrade* as predictor; iii) using both lagged measures combined; iv) both lagged success measures plus the rest of the independent variables. Furthermore, due to the difference in the structure of Bachelor degrees as compared with integrated Masters, in this model only Bachelor students were considered. Another important division concerns the type of Higher Education. It was found that universities and polytechnics have distinct distributions of scores and that these are not necessarily predicted by the same factors. Thus, for all of the specifications, both cases are analysed separately.

Table 2: Determinants of Deviation from Final Score (DevFinalScore)

	University				Polytech.	>150	<=150
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ExMScore	.013***	-	.003***	.003***	.004***	.009***	.007***
HSGrade	-	.026***	.023***	.022***	.014***	.031***	.029***
Male	-	-	-	-.002	-.052***	-.029	.045
Working	-	-	-	-.159***	.051	-.574***	-.051
Support	-	-	-	.155***	.108***	.191**	.052
Displaced	-	-	-	-.104***	-.026	-.035	-.038
PubHS	-	-	-	.092***	.089***	.013	.023
BothHS	-	-	-	.128**	.143***	.269**	-.241**
Academic	-	-	-	.210***	.210***	.343***	.251***
PSchYrs	-	-	-	-.005**	-.007***	-.007	-.007**
constant	-1.55***	-3.82***	-3.79***	-3.82***	-2.56***	-6.56***	-5.18***
N	11213	11213	11213	11213	12419	2026	5778
$R^2$	.079	.147	.151	.162#	.104#	.236#	.224#
Adjusted $R^2$	.079	.147	.151	.161#	.103#	.232#	.222#
* p-value < .10 ** p-value < .05 *** p-value < .01							
# - Mean over 20 imputations							

Looking at the first three specifications, one can notice that both lagged success measures are significant predictors of the deviation from the final score. Yet, the *HSGrade* consistently explains more of the variation, in comparison with *EMScore*: 14.7% versus 7.9% in University (8.2% versus 4.5% in Polytechnic). Looking at (3), for University, 10 more points in *EMScore* imply, on average, a final score 0.03 values further from the average, while the same variation on *HSGrade* leads to a 0.23 higher final score. This

fact, that *EMScore* loses relevance when combined in the same estimation with *HS-Grade* is partially explained by the high correlation between the two variables, which is about 65%<sup>15</sup>. Still, substituting *EMScore* by the score on Portuguese and Mathematics exams yields very similar results, adding to the robustness of this finding.

Turning now to the full specifications (eq. (4) - (7)), some results can be highlighted. The explained variability does not increase much, passing only from 15,1% to 16,2% in the university case (eq. (4)). At the same time, the coefficients for the lagged success measures remain similar. Gender is only significant in polytechnics (eq (5)), with coefficients of males equal to -0.052 in the deviation from the final score. Working while studying is only relevant for University students and seems to be negative, meaning less 0.159 values, on average. *Support* is significant and positive for both kinds, leading to an increase of 0.155 in University. Being displaced is only significant for University, and negative, meaning one decimal point less in the final score. Having been in a public school seems positive, with students from “mixed” tracks reaping the most benefits, in comparison to students that completed High School on a private school. Coming from an academic track is very significative, and similarly relevant for both kinds of Higher Education, bringing a benefit of over 2 decimal points. This is different from what would have been expected, as supposedly an academic track should be more beneficial for a student in University, relative to a student in a polytechnic. Finally, the parents average schooling years are relevant and slightly negative for University and Polytechnic. This is contrary to the literature and should be studied further.

In another specification, public universities were divided according to their Minimum Entry Grade (MEG). This can be used as a proxy for quality of the institutions.<sup>16</sup> Thus,

---

<sup>15</sup>This result could be due to the fact that we are aggregating “core” subjects (such as Portuguese and Mathematics) with all the rest, when maybe only the first would matter for the dependent variable. However, there is no complete data on the exam scores of these subjects. Mathematics is not mandatory and, in the case of the Portuguese exam, data is not available when students took the exam in the year previous to their entry in Higher Education.

<sup>16</sup>Degrees are selected by students on the basis of their quality, or reputation of quality. Then, after this choice, both

these institutions were grouped into the ones with a MEG higher than 150, and the ones with MEG lower than or equal to 150, equations (6) and (7), respectively. Some of the features seen previously are confirmed in this setting: the lagged success measures maintain its pattern; *Academic* also presents similar results, although more relevant in more demanding degrees; and *PSchYrs*'s coefficients are similar, despite the loss of significance in more selective degrees. The explanatory power of the regressions suffers a considerable increase (from 16.2% to 23.6% and 22.4%), although some variables lose significance in both groups: *PubHS* and *Displaced*. Moreover, three results should be noticed, regarding the difference between these two groups. Firstly, the loss of performance of working students seems to be concentrated only on the “better” degrees, in which working at the same time as studying implies -0.574 in the final score. This can reflect a higher difficulty of these degrees: working students may only be negatively affected by their smaller availability for coursework in these places where such work is supposed to be more demanding or time-consuming. Secondly, the same phenomenon (to a lesser extent) seems to happen with students receiving Social Support, where they only surpass their peers in the case of the degrees with a higher MEG, having almost two decimal points more, on average. This can be justified by the rules regulating the maintenance of such support.<sup>17</sup> Thus, students have an incentive to have better scores and not flunk, in order not to lose their scholarship. Still, the effect of this support may not be perfectly indentified in these specifications. This is because the regressions compare students with support with the general population, not taking into account that these students have a different (and less advantageous) background. Thus, these

---

public and private Higher Education institutions proceed to selection of their students. Public institutions do this on a centralized and determined way, through the “Concurso Nacional de Acesso”, ordering applicants by their Access Score, composed by their High School Grade and the “specific exam” score, which is determined for each degree. Thus, a more reputed degree will be more demanded and have a higher Minimum Entry Grade.

<sup>17</sup>In essential, to maintain this support, students must complete a determined number of ECTS each year and graduate in a determined deadline according to the established duration of the degree. For more detail, see Appendix C.

estimations of the coefficient may be biased downwards, meaning that the actual effect may be even stronger. Finally, there is the case of students with the mixed track, coming from public and private high schools. The results here are unclear, as it seems that these students are better performers in the first group of degrees, while performing worse in the other group. This could imply that there are some students that change the kind of schooling because they are in a poor performance situation, and those are the ones that go on to “normal” degrees, and some others that change for other reasons, which go on to “better” degrees. In the first case, the coefficient detects this “poor performance” signal, while in the latter, the coefficient could be capturing a positive “diversity” effect. In any case, these issues require further study.

The second model presented concerns the probability of graduating later than predicted<sup>18</sup>, *DelayedDegree*. This probability is predicted by a *probit* model, estimated by Maximum Likelihood. Estimated results appear in Table 3. In general, the predictive factors considered are similar, except for one: regarding *PSchYrs*, various attempts were made with inclusion of the variable, using the multiple imputation samples. However, the coefficient for this variable was insignificant in every specification, suggesting the hypothesis of no effect on Higher Education performance. Thus, the original database was used for this model, not including this variable. Moreover, no structural difference between Bachelors and Masters degrees was found in the probability of delaying graduation. Thus, estimation was proceeded including also Masters graduates. Furthermore, in all specifications, the average years taken to graduate (at the degree level), *AvgYrsTaken*, is included in order to control for degree-specific factors contributing to this probability. The presentation of results follows the same logic as before. For simplicity, only the average marginal effects are presented.

---

<sup>18</sup>This is defined as the student graduating with at least one more register than the regulated duration of the degree, by the institution.

Table 3: Determinants of Delayment of Degree (DelayedDegree) - Marginal Effects

	University				Polytech.	>150	<150
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ExMScore	-.004***	-	.000	.000	-.001***	-.002***	-.000
HSGrade	-	-.008***	-.009***	-.008***	-.005***	-.009***	-.008***
Male	-	-	-	.078***	.121***	.047***	.051***
Working	-	-	-	.036*	.024*	.185***	.028
Support	-	-	-	-.176***	-.123***	-.144***	-.209***
Displaced	-	-	-	.012	-.027***	-.012	.003
PubHS	-	-	-	-.048***	.038***	-.012	-.010
BothHS	-	-	-	-.106**	-.015	-.086**	.008
Academic	-	-	-	-.143***	-.184***	-.201***	-.220***
N	14993	14993	14993	14993	12419	3943	7038
PseudoR <sup>2</sup>	.032	.082	.082	.115	.226	.202	.144
* p-value < .10 ** p-value < .05 *** p-value < .01							

As Table 3 shows, the results are similar to the previous model, regarding past success measures. Moreover, in (3), when combining *EMScore* with *HSGrade*, the first variable is now not significant in relation to the probability of graduating late. Like before, different specifications using only core subjects instead of *EMScore* were included, and the results were, again, similar. Looking at the marginal effects, an extra 10 points in *HSGrade* meant around 25% less probability of graduating late in a university (20% less in Polytechnic). Here, contrarily to the effect on *DevFinalScore*, gender is significant, males having an 8% (12% in Polytechnic) higher probability of delaying their degree. These results are in line with the literature. It could be interesting to test whether these effects differ between areas of knowledge. *Working* is only significant at 10% in both kinds of Higher Education with an only slightly positive effect on this probability. As for Social Support, it is consistently positive for success, implying 18% less probability of having a delayed degree in University. Being displaced seems only to affect students in Polytechnic, being 2.7% less likely to delay graduation. With High School types, results seem unclear for public schools, with opposite effects on University and Polytechnic, and *BothHS* being positive for success (negative for probability of having a delayed degree) only in University. Again, students coming from an academic curricula in High

School benefit without question, adding to the robustness of previous finding.

Similarly to the analysis of *DevFinalScore*, a division by the Minimum Entry Grade was done. Here, even though some of the coefficients lose significance, the percentage of variability explained improves for both cases, although it increased much more in degrees with higher MEG (Pseudo R-squared is now 20%, coming from 11%). *Displaced* and *PubHS* lose significance in both groups. Looking again at the lagged success measures, the results are similar and add to the robustness of the found pattern. With gender, the findings are also similar, yet less relevant. *Working* loses significance in “normal degrees”, while increasing by 18% the probability of delaying graduation in the other group. As in the previous measure of success, here also it seems that difficulty of coursework influences how working students perform. Yet, although there is higher influence in marginal effect, there seems to be no influence of degree difficulty on *Support*, as its effect is felt on all specifications. Finally, *BothHS* loses significance for degrees with lower MEG, while *Academic* seems to influence both groups in a similar way, consistent with previous findings.

As mentioned before, the previous results are only representative of the population of students accepted in a specific degree and graduated in that same degree. Relating to this latter condition, which is not random, it cannot be accounted for, due to the lack of data on students who did not graduate. However, the allocation of students by institutions and degrees is also not random. More successful students in High School are allocated into institutions with a reputation for better quality. Thus, being in a “top quality degree” is in itself a measure of success in Higher Education. In fact, graduating from a more reputed and demanding institution is seen by the labour market as indicative of higher productivity, comparing to graduation from other institutions. Therefore, in order to look at what affects this distribution, OLS estimates of the

determinants of *HSGrade*, *EMScore*, *PortugueseExam* and *MathExam* are presented. The first is a component of the previously mentioned Access Score, and the other three are proxies for the other component, the “specific exam score”.

Table 4: Determinants of the components of the Access Score

	HSGrade	EMScore	PortugueseExam	MathExam
Male	-3.28***	-1.16***	-6.81***	.860
PSchYrs	.918***	1.36***	1.25***	1.39***
PubHS	-5.13***	-4.17***	-3.66***	-6.06***
BothHS	-5.63***	-9.04***	-6.65***	-10.38***
constant	143.88***	117.09***	115.07***	116.38***
N	27412	27412	21182	15892
$R^2$	.077#	.061#	.061#	.065#
* p-value <.10 ** p-value <.05 *** p-value <.01				
# - Mean over 20 imputations				

These regressions are much more limited, having only contemporaneous inputs: *Male*, *PSchYrs*, *PubHS*, *BothHS*, including also controls for Academic/Vocational curricula, and the main curricula paths of the academic track. As seen above, one can notice a negative, robust effect of being male, much more significative in the *PortugueseExam*. Thus, besides a robust result regarding later graduation, males are probably also less represented in the more demanding degrees, since *Male* has a significant and relevant influence on the components of the Access Score. Also, there is a positive effect of the average schooling years of parents. This could explain our previous findings about this variable. It could be that family background influences achievement positively in Higher Education in an indirect way, through allocation, but that there is a slight negative effect within a degree/class, due to reasons which can only be clear with a deeper research. Regarding the type of High School, there are consistent and significative positive effects of coming from a public high school, in the general specifications. Yet, these effects are, as said before, conditional on the existing distribution of students, which is in itself determined by the Access Score, in turn affected by the type of high school where

the students come from. This indirect effect shows to be negative in our regressions on the determinants of the components of the Access Score. Despite this, one must realize that these OLS estimations include only contemporaneous inputs. Thus, the total effect on success of coming from a public high school remains undetermined, even though the conditional effect is positive and relevant. Still, one must account for the restricted sample of the population considered (High School students who went to Higher Education and graduated) and take notice of the limited explaining power of these specification, which amounts to an R-squared of 8%, at most.

## 6 Conclusion

The main and more robust finding provided by the evidence presented is related to the *EMScore* and *HSGrade*. It seems that the high school’s internal grade is much more important for success within a class, given the existing distribution of students through degrees and departments<sup>19</sup>. This is in line with the finding of Alves (2014), for students of Economics and Management. Moreover, this does not mean that the National Exams lose importance, as it seems obvious that internal grades are given by teachers taking into account the expected performance of students in those exams, and that students study knowing that they will have final exams.

The most surprising results are, however, related to the effect of parent’s level of schooling. The findings show that effects are, at most, slightly negative, contrarily to the existing literature. These effects are, yet, conditional on the distribution, and the effects on the components of the Access Score for the “preliminar” estimation presented on

---

<sup>19</sup> Still, one must account for the problems of *EMScore*, as it is defined: it comprises more exams than the ones used for the Access Score and it is not guaranteed that students did not use for the Access Score an exam done on the previous year.



these show a clear and relevant positive effect. Nevertheless, these findings give some strenght to an hypothesis that the effects of the schooling of parents may have been already entirely absorbed during the previous stages of formal education.

Moreover, it is worth noting the interesting results regarding the differences in coefficients of various factors according to type of Higher Education and quality of the institutions attended by students. For example, it seems that working students are more likely to have worse scores and delay their graduation, but only in degrees where the Minimum Entry Grade is larger than 150, with reputation of being most demanding. This research comes with many limitations. The most important ones refer to the sample being used. In fact, as explained, the sample corresponds to part of the students that graduated in 11/12. Yet, this is by no means a cohort. The ideal database would include all students that entered in Higher Education in a given year, and, from there, find what determines their success. Furthermore, an ideal research on the determinants of achievement would have to include some measure of success in the labour market after graduation, as this is the ultimate goal of Higher Education. This, however is very difficult, not only because of sheer availability of data, but also because of confidentiality issues. Moreover, in this analysis, the fact that these databases were not open for the public also implied that they were also not necessarily adapted for the use of the researcher. The database itself was created ad hoc by the author from at least five different sources of information, leaving room for possible mistakes or errors. The recommendation is that, for the future, data is made available for researchers, already compliant with data protection laws and requirements. In this way, not only will more research be done, but also, due to communication from researchers to the authorities, the data available will be much more adequate for research. Finally, it will open the door for results to be checked by other independent researchers.

## References

- [1] Alves, Dino. 2014. Determinants of Success of Nova SBE's undergraduate students. Master's thesis, Nova School of Business and Economics.
- [2] Amann, Roland. 2005. *The Effects of Employment on Time-to-Degree in Higher Education: Does the Type of Employment Matter?*. Working Papers of the Research Group Heterogenous Labor 05-06. Research Group Heterogeneous Labor, University of Konstanz/ZEW Mannheim.
- [3] Becker, Gary S. 2009. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.
- [4] Cunha, Mariana, and Vera Rocha. 2012. *On the Efficiency of Public Higher Education Institutions in Portugal: An Exploratory Study*. FEP Working Paper. Universidade do Porto, Faculdade de Economia do Porto.
- [5] DesJardins, Stephen L., Dennis A. Ahlburg, and Brian Patrick McCall. 2002. "A Temporal Investigation of Factors Related to Timely Degree Completion." *The Journal of Higher Education* 73 (5): 555–81.
- [6] Evans, W. N., and R. M. Schwab. 1995. "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *The Quarterly Journal of Economics* 110 (4): 941–74.
- [7] Feinstein, L., and J. Symons. 1999. "Attainment in Secondary School." *Oxford Economic Papers* 51 (2): 300–321.
- [8] Freeman, P. R. 1970. "A Multivariate Study of Students' Performance in University Examinations." *Journal of the Royal Statistical Society. Series A (General)* 133 (1): 38.
- [9] Hanushek, Eric A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *The Journal of Human Resources* 14 (3): 351.
- [10] Horowitz, John B., and Lee Spector. 2005. "Is There a Difference between Private and Public Education on College Performance?" *Economics of Education Review* 24 (2): 189–95.
- [11] Krueger, Alan, and Mikael Lindahl. 2000. *Education for Growth: Why and For Whom?*. w7591. Cambridge, MA: National Bureau of Economic Research.

- [12] Marcenaro, Oscar, and Lucía Navarro. 2007. “El éxito en la universidad: una aproximación cuantitativa,” *Revista de Economía Aplicada*, 15:5-40
- [13] McNabb, Robert, Sarmistha Pal, and Peter Sloane. 2002. “Gender Differences in Educational Attainment: The Case of University Students in England and Wales.” *Economica* 69 (275): 481–503.
- [14] Messer, Dolores, and Stefan C. Wolter. 2010. “Time-to-degree and the Business Cycle.” *Education Economics* 18 (1): 111–23.
- [15] OECD. 2014. Education at a Glance 2014. Education at a Glance. OECD Publishing.
- [16] Plug, Erik, and Wim P. Vijverberg. 2001. *Schooling, Family Background, and Adoption: Is It Nature or Is It Nurture?*. IZA Discussion Paper 247. Institute for the Study of Labor (IZA).
- [17] Siegfried, John J., and Wendy A. Stock. 2001. “So You Want to Earn a Ph.D. in Economics?: How Long Do You Think It Will Take?” *The Journal of Human Resources* 36 (2): 364.
- [18] Simkins, Charles. 2005. *The Determinants of Educational Attainment*. Working Paper 31. Economic Research Southern Africa.
- [19] Smith, Jeremy, and Robin Naylor. 2001. “Determinants of Degree Performance in UK Universities: A Statistical Analysis of the 1993 Student Cohort.” *Oxford Bulletin of Economics and Statistics* 63 (1): 29–60.
- [20] Todd, Petra E., and Kenneth I. Wolpin. 2003. “On The Specification and Estimation of The Production Function for Cognitive Achievement.” *Economic Journal* 113 (485).
- [21] Universidade Do Porto. 2012. *Percurso de Estudantes Admitidos* 2008, v 19dez. 2014. Accessed August 27.

## Appendix

### Appendix A - Tasks performed at the Directorate-General of Statistics of the Ministry of Education

The internship at the Ministry of Education was part of the work done for the Masters in Economics final Work Project. In particular, the topic studied was previously determined by the Directorate-General. In this context, tasks performed during the time at the ministry were partially motivated by the topic of the thesis and, most of the times, instrumental for the accomplishment of this paper's goals. Nevertheless, some other tasks were not directly related with the topic of research. Still, all the work done throughout the six months of internship was useful for the DGEEC, either directly, or indirectly through its usefulness for the investigation. Next, a brief list and description of some of the main work done is presented:

- Training in SQL Server and Microsoft Access, joined with an extensive browsing of all data at the Ministry that would potentially be used for this paper.
- Completeness analysis of the variables available at the Ministry database, with focus on the ones potentially to be used on the investigation; Production of various reports on this issues, with highlight to a report on the completeness of the variables related to the Schooling of parents (summary of report presented in Appendix B).
- Reliability analysis of a set of variables whose characteristics allowed for validation processes to be done, with focus on those variables for which there were already some suspicions of poor consistency. Highlight on the reports regarding “Number of Registrations” and “ECTS” (summary of these reports presented in Appendix B).
- Construction of the database to be used for the present paper, with collection of

data from (mainly) 5 different sources (most of these internal to the DGEEC), pairing of corresponding observations in different sources and forcing of various conditions on the database to ensure reliability and homogeneity of the database.

These documents should be useful, not only for research, but also for the statistical activity of the Directorate, as they detected failures on the reporting of variables (typical mistakes by reporting entities) and gave suggestions for the correction of these problems. Afterwards, new proceedings for next year's inquiry have been decided to be put in place, as to reduce some of the problems reported. On the basis of this analysis, the final list of variables was selected for the construction of the database, together with some conditions that allowed for both the correct identification of the population and the consistency and reliability of data.

On the basis of this analysis, the final list of variables was selected for the construction of the database, together with some conditions that allowed for both the correct identification of the population and the consistency and reliability of data. In the next section, an example of reliability analysis will be presented, together with a description of the necessary steps for the transformation of the database.

Also regarding the selection of data, the decisions taken had much to do with availability: only the cohort of students finishing their degree in 2012 could be associated and linked to their own information when signing up for university, due to issues in the individual registration of personal identification data. Also, out of these, only the ones that took the National Exams after (and including) 2006 could be used. Following this, an evaluation of specific variables was undertaken, looking at completeness and, in some cases, reliability, through the use of consistency conditions relating different variables.

## **Appendix B - Summary of Reports produced at the DGEEC**

### **Report 1: Completeness of variables related to the Schooling of parents**

This report evaluates two variables that state the Schooling of the student’s father and mother through various categories (e.g.: Bachelor’s degree, Master’s degree, etc.). One of these categories is labelled “Unknown” and, as this information comes from student Registrations, it means that, at the time of the registration, the student was not aware of the level of schooling of her/his father/mother.

The document presents the statistics for the registered students in 2011/12, showing that, for Bachelor and Integrated Masters students, the percentage of “Unknown” observations was around 26% for both father and mother’s schooling, and about 23% for “Unknown” on both variables on the same observation.

This poses a problem because this “Unknown” category is correlated at the same time with the success measures and (most probably) with the actual Schooling of parents. This first correlation is shown by the distribution of students by final scores and by time to degree: in both cases the distribution of students with both “Unknown” parent schooling is worse than the distribution for the total of the population.

Moreover, looking at intra-department distributions, this phenomenon is still seen. Thus, the issue is not related with greater percentage of “Unknown” schooling in faculties where time to degree is longer and final scores are worse.

Some possible explanations are presented, and suggestions are provided for the improvement of reporting by faculties, which involve a less confusing categorization of schooling intervals. Finally, a ranking of the faculties by percentage of incompleteness on these two variables is presented. This allows for a greater control by the part of the DGEEC, since reporting efficiency varies widely among different institutions, and greater pressure on the part of the Ministry could prove effective in some cases.

## **Report 2: Reliability of the variable indicating the number of registrations until graduation**

This analysis relates to the consistency and reliability of a variable that reports the number of registrations until graduation present on the Graduates table. Basically, if this variable is reliable, and using only students that have not interrupted their academic path, the number of registrations is equivalent to the number of years taken to graduate. However, there was a suspicion on the consistency of this variable.

In order to validate this reporting, this table of Graduates was linked by coincident observations to the table of Registrations. Basically, students that were found in the Graduates table of 2011/12 were also found in the Registrations table of that same year. Also, this second table also presents a “number of registrations” variable, which corresponds to the number of registrations in that same degree and faculty until the present one. Thus, students found on the same degree and university on both tables should present consistent values (for example, a student registering for the third time and finishing that year should have 2 registrations in the Registrations table and 3 registrations in the Graduates table). The conclusion was that almost 20% of these observations were not consistent, and, in this sense, unreliable.

After, an analysis on the 70 institutions with more than 15% of inconsistencies was made, in an attempt to understand the reason for some of these problems. Still, although a significant part of these were found to be (most likely) reporting errors with a distinguishable pattern, no definite conclusions were taken from this specific analysis. Moreover, a similar validation was done for two consecutive years of Registrations, using coincident students, which showed similar results.

Finally, a ranking of faculties according to the percentage of these inconsistencies was also provided to the DGEEC, so as to allow for a deeper analysis of this problem in the future.

### **Report 3: Reliability of the ECTS variable**

In this particular analysis, the focus was on a variable reporting the course credits (known in Europe as ECTS) previously done in that degree and university, reported on the Registrations table. Thus, the variable should report, in the yearly registration at the university, the number of credits already completed by the student. If reliable, this variable could also be used as a measure of relative success with inclusion of students which have not yet finished Higher Education but are still pursuing their degree.

Again, this variable also presented some unrealistic values, which led to the need of further analysis. The analysis was similar to the one done with the “number of registrations”, in the fact that also the data on coinciding students from two consecutive years was used. In a normal path, a student completes 60 ECTS each year, completing a 3-year degree with 180 ECTS, a 4-year degree with 240 ECTS, etc. Moreover, there are restrictions to completing more than 60 ECTS per year: it is possible, but more than 75 ECTS per year is very rare (a realistic percentage should not be over 2-3%). Thus, an analysis of the differential of credits completed in successive years was made. The results showed that, not only there were 15% of students with more than 75 credits completed in one year (over 8% with more than 90 credits), but also that almost 5% of students had a decrease in the number of ECTS completed, which is technically not possible. Yet, these “excessive credit” completion could be justified by completion of Work Projects or Dissertations, which represent a large chunk of credits by themselves, while at the same time a student is completing subjects. Still, looking at the same differential only for the progress between the first and the second registration of the student, similar percentages are found.

Thus, this variable was deemed unreliable for statistical use and, in particular, for this research. Similarly, a ranking of faculties by percentage of these “unrealistic values” was provided to the DGEEC, in order to improve the quality of reporting of statistical information.



## **Appendix C - The Portuguese Higher Education System**

The Portuguese Higher Education System is a set of various institutions that provide post-secondary Higher Education, integrated in the Bologna Zone. As such, among other types of qualification, first cycle (bachelor), second cycle (masters) and third cycle (PhD) formation is provided. In Portugal, first and second cycle are offered separately or in a joint longer degree, called an Integrated Masters, combining both Bachelor and Masters degree.

Moreover, coming from an already existing framework, Higher Education institutions are divided into Universities and Polytechnicals. While both of them offer Bachelor and Masters degrees, only universities can provide Integrated Masters and PhD's. Also, universities are required to pursue research, while Polytechnicals are not. Traditionally, Polytechnicals have specialized in less theoretical and more industry-related fields, although they are not necessarily restricted to those subjects. Furthermore, these institutions tend also to be more spread and present in less populated areas, while Universities are only found in relatively more populated cities.

There is another distinction that regards property of the institutions: private or public. Public institutions have a reputation for being better and more demanding in most fields, although there are some private institutions which are also of high quality.

Access to Higher Education varies between public and private institutions. Allocation to Public Higher Education is done through a centralized proceeding through which students choose 6 options (each option is one degree in one institution) with a specific order of preference. Students are then ordered for each of the degrees according to their Access Score. The Access Score is determined both by the Final High School Grade (50% to 65%) and by a specific National Exam determined by the institution (35% to 50%). A series of iterations is then computed to allocate students according to their Access Scores and ordered preferences. Private institutions also use National Exams and High School Grades, but they are free to determine their entry criteria and it varies

widely among institutions.

Tuitions in Portuguese Higher education are relatively cheap in public institutions (around 1000 euros per year), and usually significantly more expensive in private faculties. Still, even for those that cannot afford such costs, there are Social Support Scholarships, for which students apply, based on family income. Moreover, to maintain these scholarships, students must complete at least 60% of the credits they have signed up for in the previous year, and they must finish the degree only up to one year later than what is determined (2 years if the degree exceeds 3 years of duration).

## **Appendix D - Multiple Imputation**

Due to the question of the incompleteness of the variables relating to parent Schooling, and taking into account the importance of this factor in the Economics of Education literature, it was found prudent to try not to drop any observations because of this factor. Thus, a method of Multiple Imputation was used to impute values on the PSchYrs (Average Years of Schooling of both parents) variable. A series of 20 imputations was made on this variable using linear regression with a wide set of available regressors, all of these with no missing values themselves. The specific way in which it was made followed the proceeding of the Stata Multiple Imputation package.